Week 6 : Descriptive Statistics (Part2)

pending tasks	
type	

Measure of centrality and spread

Learning objectives:

1. Why do we need measures of spread and centrality?
2. What are different measures of centrality?
3. What are the characteristics of these measures ?
4. What do measures of centrality look like for different types of distributions?
5. How to compute measures from histograms?
6. What is the effect of transformations on these measures?

Why do we need measure of centrality and spread?

Large amounts of data can be succinctly summarized using these measures.

Summary statistics are used for quantitative data.

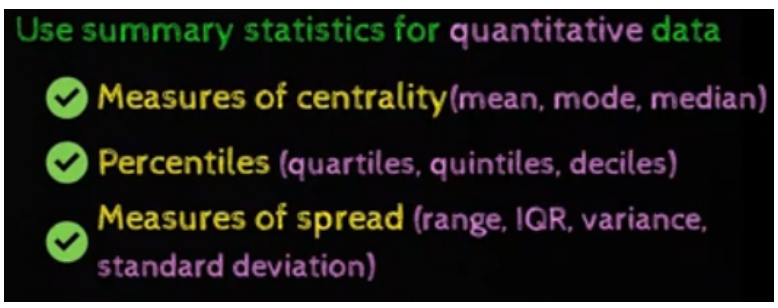


Fig.1 use of summary statistics

Different measures of centrality

Notations used - (t = 1:41)

def mean - (t= 1:23 , t = 12:56)

def median - (t = 5:45)

def mode - (t = 10:54)

Measures of centrality are used to depict the typical value of an attribute in the dataset.

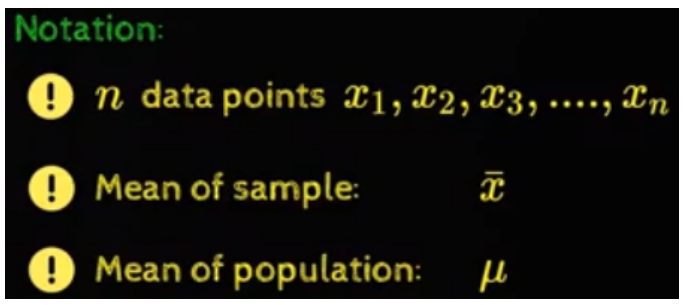


Fig.2 Notations used

- **Mean** is given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **Median** is the value that appears in the centre of sorted data. There are equal number of elements on either side of the median.
- When n(total number of elements) is odd the median is the value at the central location (the midpoint).

- When n is even there cannot be a single midpoint. The data will have two midpoints. Thus, the median is the average of two points.

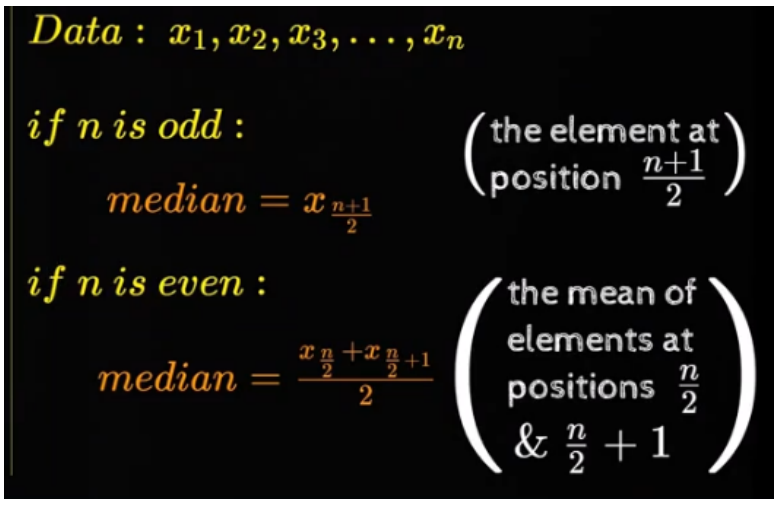


Fig.3 Formulae to calculate median.

- **Mode** is the most frequently occurring value in the dataset. The data can have more than single mode and is called a multimodal distribution. If there are two modes, it is called a bimodal distribution.
- If every value occurs only once, this implies there is no mode in the data.

Characteristics of measures of centrality

def deviation - (t = 0:35)

- The mean is known as the **center of gravity**. The deviation of a point from the mean is defined as the difference between this point and the mean.

$$deviation = x_i - \bar{x}$$

- **The sum of deviation of all points from mean is 0.**

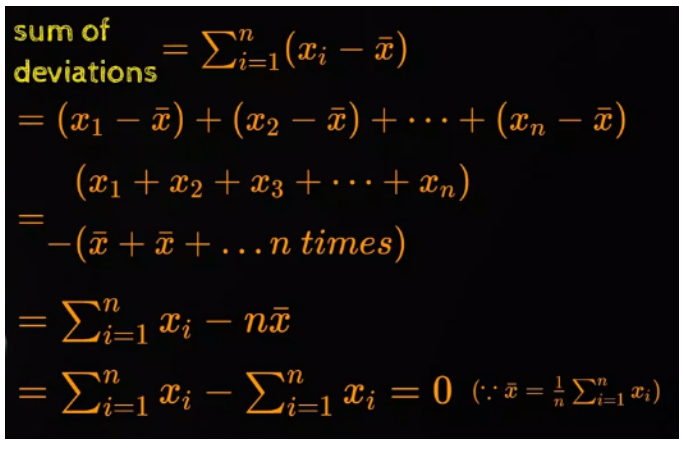
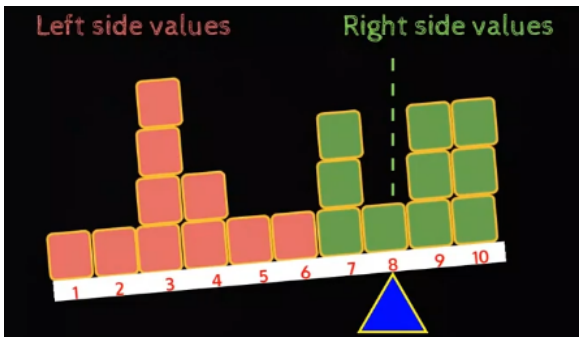
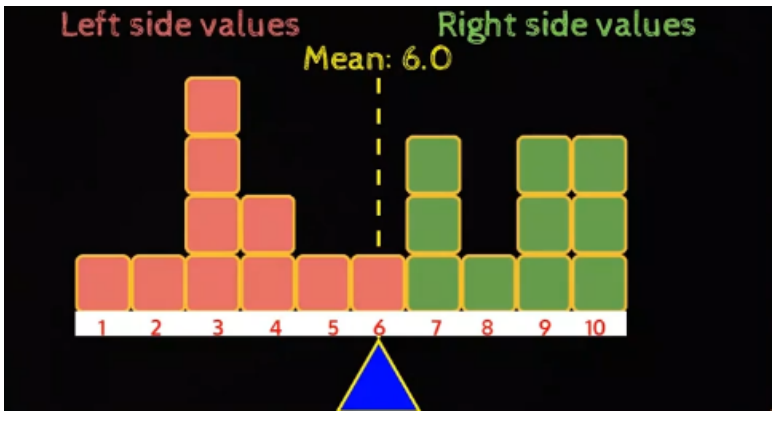


Fig.4 Proof for sum of deviations = 0

- **Physical interpretation of the result:**

Consider the number line as a seesaw and the data points as weights on the seesaw. weight s are proportional to deviations from x .





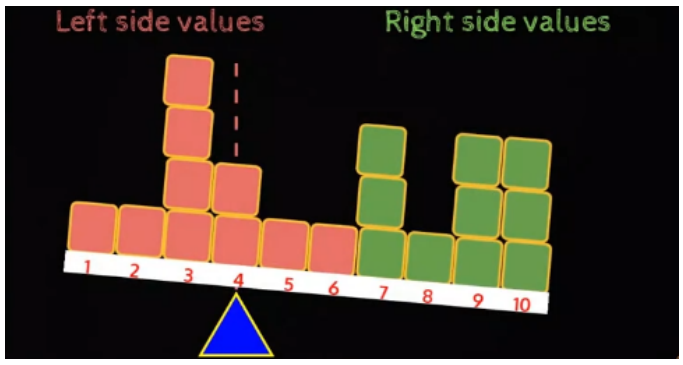


Fig.5 Physical interpretation of the results.

Mean balances the weights on either side of the seesaw. Thus, mean is called the centre of gravity.

Sensitivity of measures of centrality of outliers

def outlier - (t = 0:47)

def trimmed mean - (t = 7:10)

- Outlier is a point away from other values in the data.
- Consistency is reflected in the median and not the mean. Median depicts the value under which 50% of the data lies and is not affected much by the presence of an outlier. Mean is affected by outliers as the outlier significantly contributes to the numerator.
- Thus, to account for the sensitivity to the outliers, it is advised to compute trimmed mean. It is computed by dropping k extreme elements (same number of elements) from either side of the sorted data.

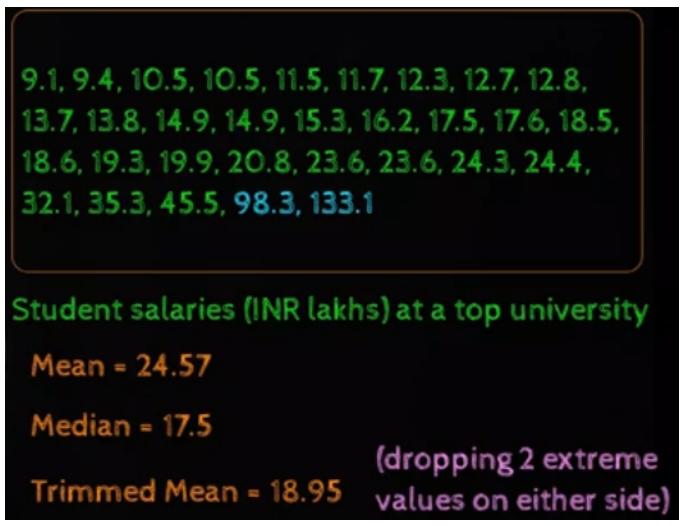


Fig.6 Example depicting how outliers affect the calculation and use of trimmed mean.

In Fig.6 the mean shows a more optimistic picture of the data whereas the median shows a more realistic picture.

- The **mode is** not sensitive to outliers, unless the mode itself is the outlier.

What do measures of centrality look like for different types of distributions?

Perfectly (unimodal) symmetric distribution

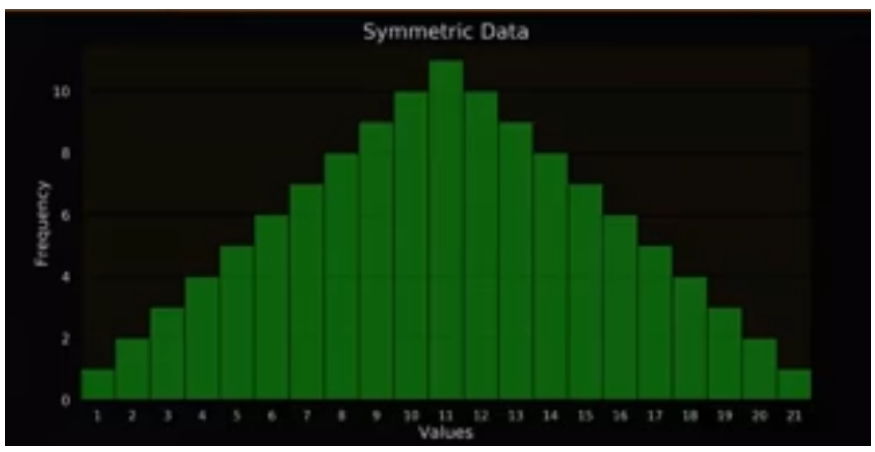


Fig.7 Perfectly symmetric distribution.

- In a perfectly symmetric distribution with central location x, there exists a (x-i) for every corresponding (x+i)th element in the data.

- **Mean = median = mode**
- Mode corresponds to the tallest bar.
- Median also corresponds to the tallest bar with an equal number of elements on either side.
- The positive deviations cancel out the negative deviations in the data. Therefore, the mean is also the central point.

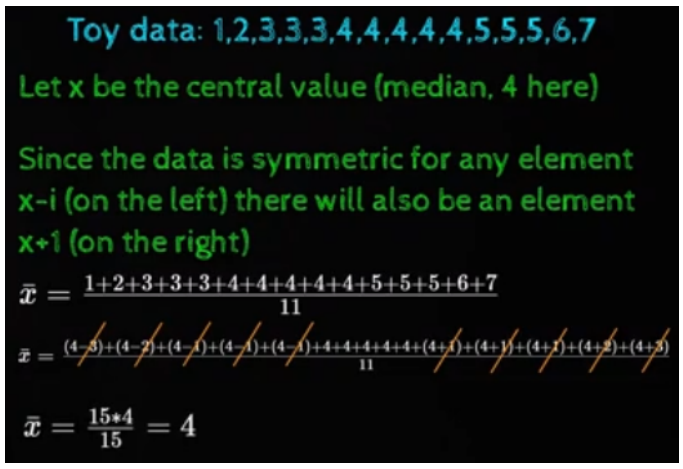


Fig.8 Informal proof for mean.

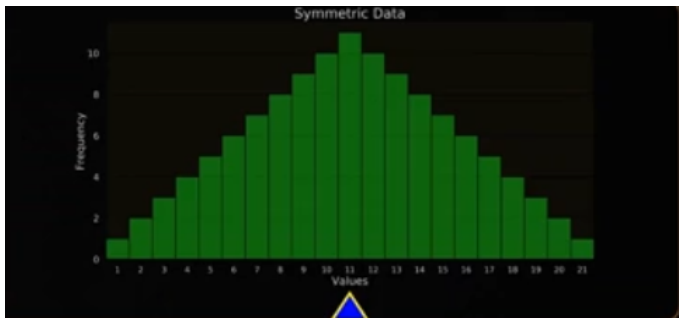


Fig.9 centre of gravity picture of mean.

It is inferred that the seesaw is balanced only when the fulcrum is placed at the median.

Bimodal symmetric distributions

> 🗒️ **Mean = median != Mode.**

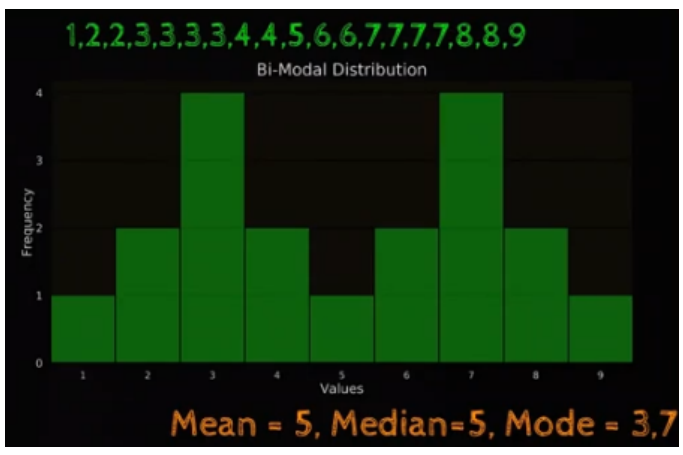


Fig.10 Example of bimodal symmetric distribution.

- In this distribution there are two modes.
- Similarly, for multimodal symmetric distribution mode is different but the mean and the median are equal.

Skewed Distributions (unimodal)

- **Left-skewed** distribution has a long tail to the left (**negative-skewed** as skew is towards the negative side of the number line) .

> 🗒️ **Mean < Median < Mode**

- The mean is towards the right (balances the heavy sided right with the light left).

- Mode will be the tallest bar.
- Median would be towards the right as there are more elements towards the right. Median represents the halfway point of the data.

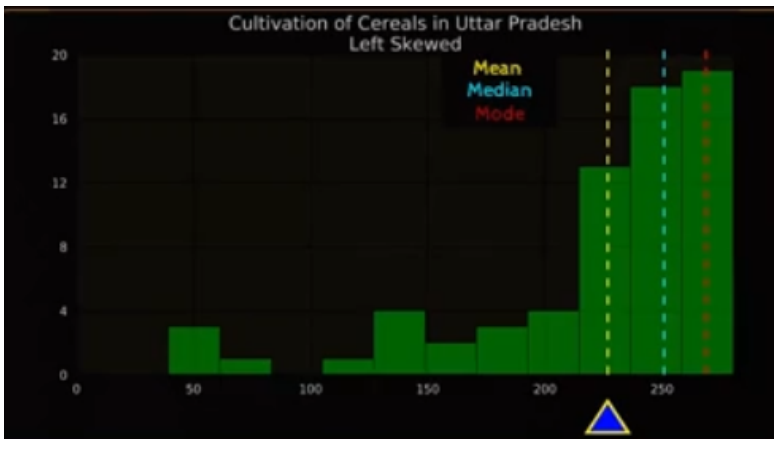


Fig.11 Left skewed data.

- **Right-skewed** distribution has a long tail to the right (**positive-skewed** as skew is towards the positive side of the number line) .
- The mean would be towards the left to balance out the weights. Median would be lesser than the mean in general, thus towards the left. Mode will also be towards the left.

> **Mean > Median > Mode**

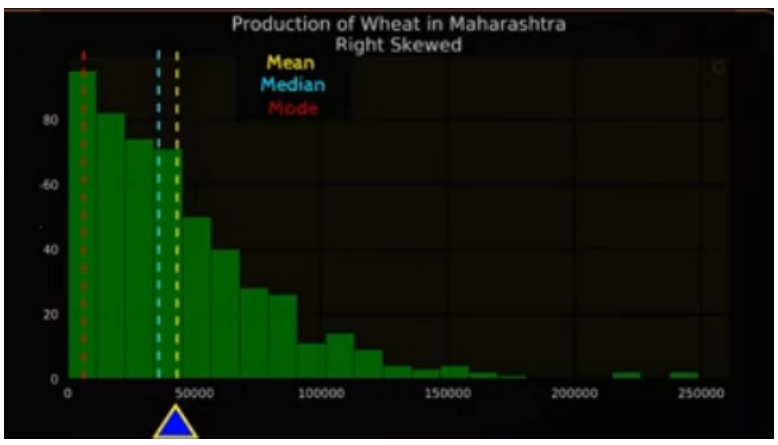


Fig.12 Right skewed data.

- This is true for perfectly left and right skewed distributions. But may not always hold.
- In skews, the mode can lie at anypoint in the opposite direction, need not always be the extreme opposite point. Eg, where there are heavy tails.
- In the fig.12 a. mean > median and fig.12 b mean< median.

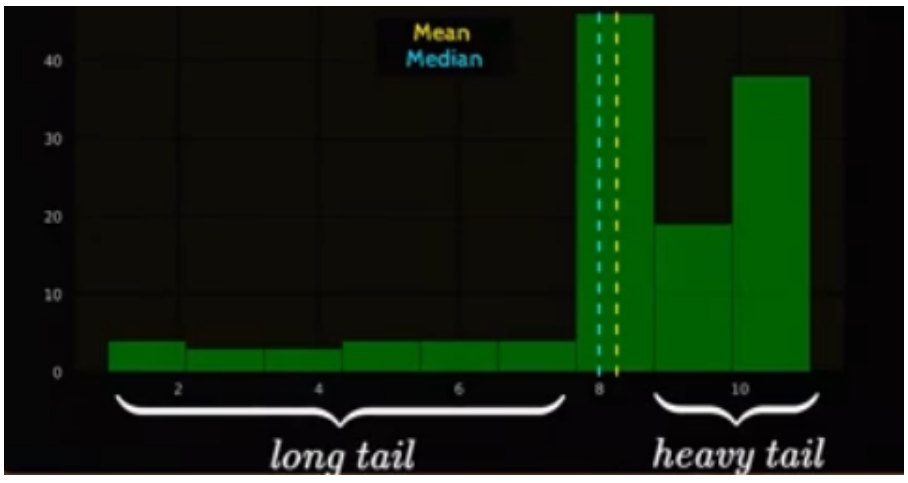
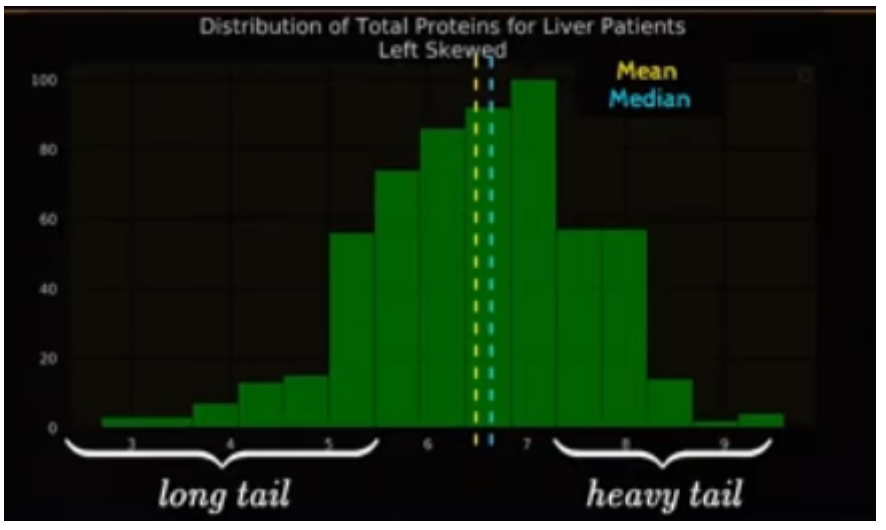


Fig.12a,b Left skewed distribution with heavy right tail.



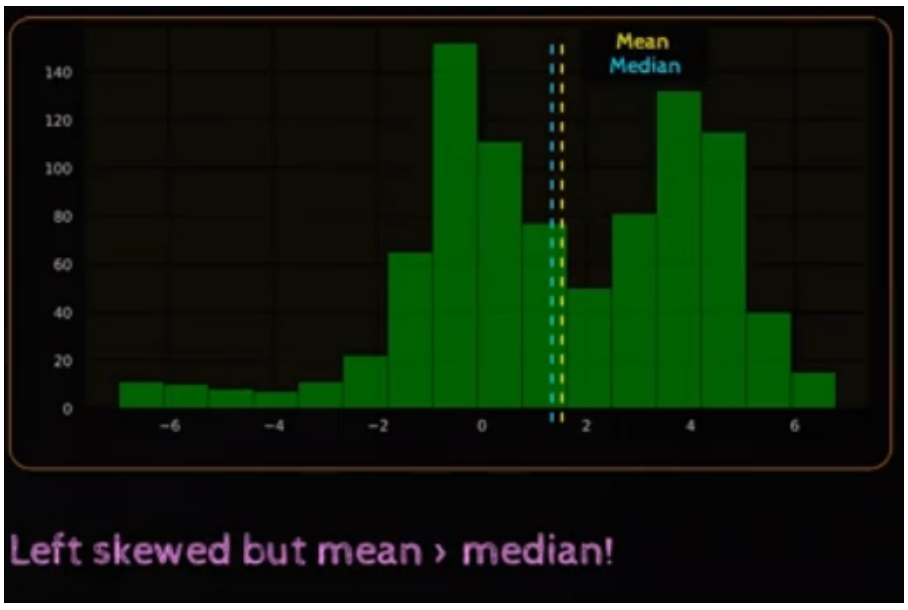


Fig.12**c. left skewed bimodal distribution**

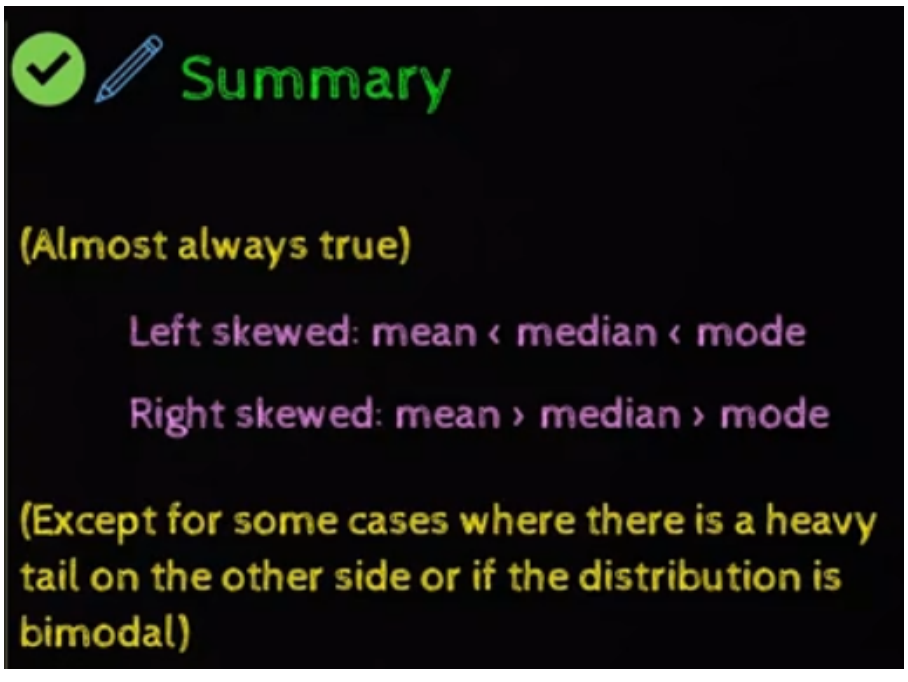


Fig.12d. summary

- **Skewed Distribution (bimodal) ,** the generic condition fails.
- **Very long tail** will not violate the generic condition.

Compute median from a histogram

- Obtain the frequency table equivalent to the histogram.
- Procedure:
  1. Compute the cumulative frequency, this accounts for the data points seen so far.
  2. Computer central location.
  3. Find the interval containing the centre. Thus, the interval containing the median can be found,beyond this there is no information available to calculate the median.
  4. Estimate the = midpoint of the interval. This is the best guess made.
- If the intervals are bigger, the absolute error might be large but the relative error may be small. Such an estimation is acceptable.

Compute mean from a histogram

- Draw the histogram equivalent frequency table.
- Procedure:
  1. Compute the midpoint of each interval
  2. Multiply the midpoint by the frequency of the interval.
  3. Sum up the resulting product for all the intervals.
  4. Divide by the number of data points.

Intuition behind the procedure:

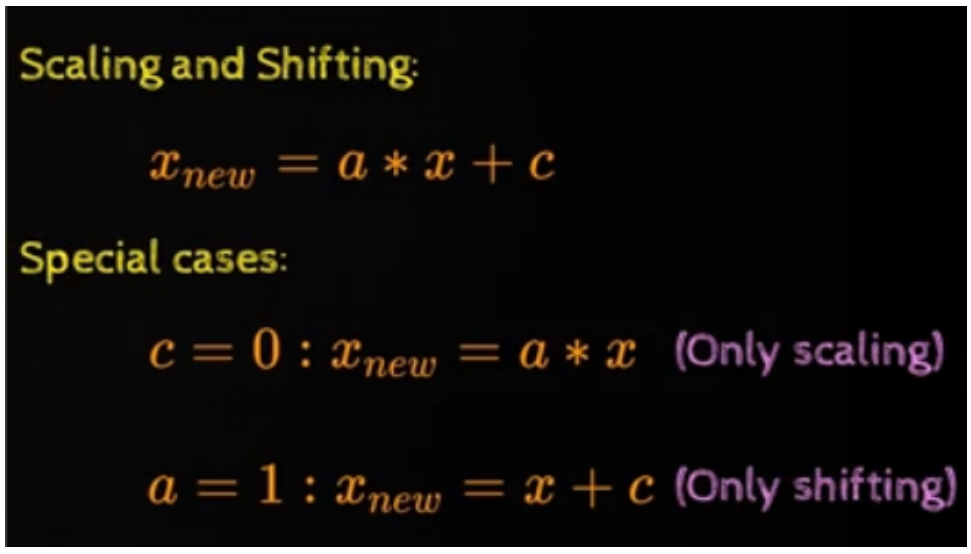
- Mean is the sum of all the elements divided by the total elements.
- The sum in the numerator can be written as the sum of elements in the corresponding intervals.
- In each interval, there is a sum of ‘x’ elements and it can be assumed that each element is equal to the midpoint. Thus, the most of the effects of underestimated elements would be cancelled out by the overestimated elements.

Compute mode from a histogram

- Mode may not be in the interval with the highest frequency.
- If the bin size (class interval) is greater than 1, it is not possible to compute the mode from a histogram.
- If the interval is of size 1, mode is the number with the highest frequency bar.

Effect of transformation on the measures of centrality

Transformations include scaling and shifting eg. Fahrenheit to Celsius.



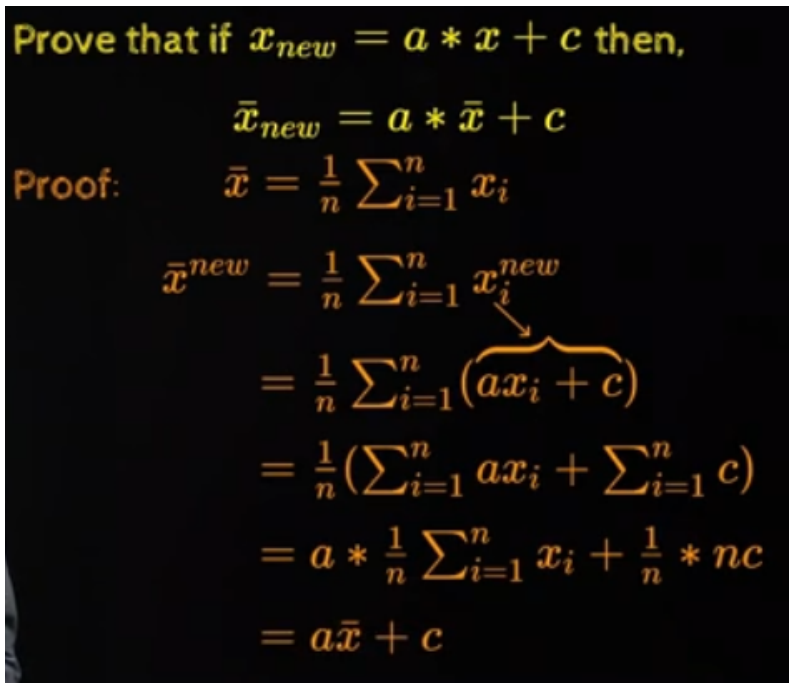
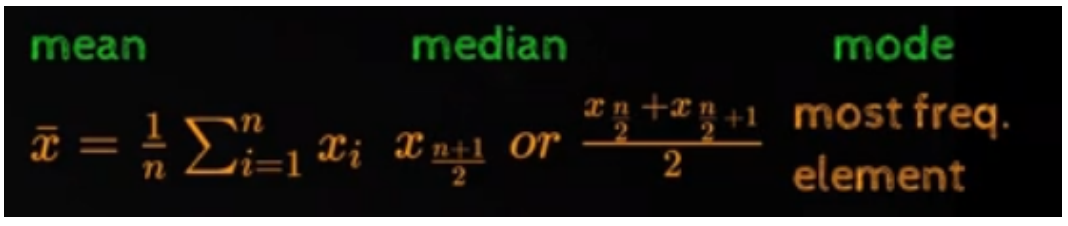


Fig.13 a.Summary of scaling and shifting , b. effect of transformation on mean

- **Effect of transformation on mean :** Mean gets scaled and shifted by the same transformation.
- **Effect of transformation on median:** The relative ordering of the elements will not change. All elements are scaled and shifted by the same amount. Thus, the median is transformed by the same amount.
- **Effect of transformation on mode :** new mode is transformation of previous mode

Summary



- Mean is sensitive to outliers, median is not.
- Trimmed mean is calculated when outliers are present in the data.
- Mean can be called the center of gravity of the data, here the deviations on the left side are equal to the deviations on the right side.
- In general for histograms with
  - Left skew : Mean < Median < Mode
  - Right skew : Mean > Median > Mode
- Mean and median can be approximately computed from a histogram. Mode can be calculated only when the interval size is 1.
- Effect of transformations on data is reflected as such on mean, median and mode (measures of centrality) i.e same transformation is to be applied to calculate the new measure .

MCQ : Week 6

1. The arithmetic mean of a given dataset is 12 and there are 20 observations. What is the sum of all observations?
   1. 8
   2. 32
   3. **240**
   4. 1.667

2. Graphs which present the values on the horizontal axis and the number of times this occurs on the vertical axis are known as
    1. Bar graph
    2. Scatter graph
    3. Box-plot
    4. **Frequency distribution**
3. When most of the scores cluster around lower end of the scale, the distribution observed is
    1. normal
    2. bi-modal
    3. **positively skewed**
    4. negatively skewed
4. When most of the scores cluster around higher end of the scale, the distribution observed is
    1. **positively skewed**
    2. **negatively skewed**
    3. heavy tailed
    4. uniform
5. Which measure is the most unreliable indicator of central tendency if the data is skewed?
    1. Median.
    2. Mode.
    3. **Mean.**
    4. None of the above.
6. Scores that differ greatly from the measure of centrality are called
    1. **Outliers**
    2. **Extreme scores**
    3. The best scores
    4. Non preprocessed scores
7. Trimming the extreme values at the left end of the dataset has the effect of
    1. Lowering the mean
    2. **Raising the mean**
    3. No effect
8. The suitable measure of centrality for average shoe size of children is
    1. Mean
    2. Median
    3. **Mode**
    4. All of the above.